Linguistic Fingerprints of Pro-vaccination and Anti-vaccination Writings

Rebecca A. Stachowicz[1]

[1] Montclair State University, Montclair NJ 07043, USA
stachowiczr1@gmail.com

Abstract. Vaccination hesitancy has been gaining public attention as a global health threat. With the echo chamber effect of social media and the prevalence of misinformation, it is becoming more important to understand all aspects of anti-vaccination attitudes, especially when facts to the contrary can solidify rather than change beliefs. Starting with the question of what might drive anti- or pro-vaccination views, this paper describes a new, balanced corpus of vaccination writings. In order to gather all the linguistic signals, corpus analysis as well as feature selection and classification tasks are used to explore themes and motivations that define each class. Our results reveal that anti-vaccination writings are typically less formal. Results also indicate the possibility that the authors of such writings are processing trauma. These findings suggest that future health promotion efforts should make attempts not to talk down to individuals and should stem from a place of understanding.

Keywords: Language, Vaccination, Health, Autism, LIWC

1 Introduction

Vaccination has been gaining recognition worldwide as a global health priority. The World Health Organization's 2019 list of ten threats to global health includes vaccine hesitancy [1]. With the echo chamber effect of social media and the prevalence of misinformation, it is becoming more important to understand all aspects of anti-vaccination attitudes, especially when facts to the contrary can solidify rather than change beliefs [2-4][19].

Anti-vaccination sentiment is nothing new. Resistance to vaccination became organized when The Vaccination Act of 1853 required Edward Jenner's smallpox vaccine by law in England. A group was formed called the Anti-Vaccination League.

Years later, Wakefield [5] claimed a link between the measles, mumps, rubella (MMR) vaccine and autism. Spurred on by this fraudulent study, anti-vaccination sentiment has grown since. Occurrence of measles, which is preventable by vaccine, increased 30% in 2017, and, "countries close to eliminating the disease have seen a resurgence" [1].

According to the Centers for Disease Control (CDC) no link exists between vaccines and autism (2019). The British Journal, Lancet, the original publisher of the

infamous study, has since retracted Wakefield [5][7]. There has also been an abundance of research conducted since that disproves the spurious link between vaccines and autism [8-11]. However, due to media attention in the UK, US, and worldwide, in addition to celebrity endorsement, Wakefield's study is not universally considered to be misinformation, and many parents still see vaccination as a personal choice. [12-13].

The idea that there is a connection between vaccines and autism has taken root in many communities, and it has grown into a movement. Just like the Anti-Vaccination League of 1853, people today, often labeled *anti-vaxxers*, are refusing to vaccinate their children. The total number of people infected with measles in 2018 was three times the number in 2017 [1], and the percentage of children under 2 years old in the United States who haven't received any vaccinations has quadrupled in the last 17 years [14]. Central to the anti-vaccination movement are a series of core beliefs, among them: vaccines contain harmful substances, vaccination is a personal choice, and vaccines cause autism in children. The latter two are especially harmful given that "herd immunity" is necessary to support individuals who cannot be vaccinated and how autism, and autistic people are often framed in anti-vaccination narratives.

Given the dangers of this global crisis, changing anti-vaccination attitudes has become a public health issue. Research has shown that confirmation bias can often thwart efforts to convince parents to vaccinate using facts [15][19]. Therefore, new and different approaches to this complex problem are not only welcome, but also necessary in the current climate.

1.1 Background

As facts continue to fall short, it is imperative that we explore other methods to ensure vaccination coverage. Previous research has focused on creating a profile of individuals with anti-vaccination attitudes [12][16], and previous studies have solidified the assumption that users of social media primarily consume information that reinforces their own beliefs [4] [15]. Attitudes towards autism and vaccination are also reinforced through other online content [17]. Terms, themes, and figures that define the vaccination debate have also been identified [18]. Language use has also been compared across these insular groups, but until now, this research has focused only on social media platforms within the limited context of comment threads, shares, and reposts [3] [17].

Research has also come to light describing the often stigmatizing nature of labeling neuro-divergent individuals [20]. While the validity of person-first language is still debated among the Autism community [21], understanding the kinds of labels and language used in both the anti-vaccination and pro-vaccination communities provides insight into each group's perception of autistic individuals. This may prove essential to public health interventions especially when autistic children and their siblings are undervaccinated [22].

1.2 Current Study

In this paper we aim to investigate the language used in web-based writings outside of social media and across genres on the subject of vaccination. This corpus analysis, as well as feature selection and classification tasks are designed to provide a tool to better understand anti-vaccination writings and the psychological aspects of the language surrounding this movement with the ultimate aim of aiding future health promotion efforts.

The goals of the study were twofold. Part 1 explored terms and Linguistic Inquiry and Word Count (LIWC) features. LIWC is a text analysis program that evaluates the degree to which various types of words are used in a piece of writing. The results of LIWC analysis provide insight into the psychological meaning behind specific language choices [23]. The hierarchical structure of LIWC enabled us to hone in on specific categories of words. To illustrate, the LIWC category Relativity includes words such as, area, bend, exit, and stop, and one of its subcategories Space includes the language items, down, in and thin among others. These LIWC features and other terms in this corpus were analyzed for frequency as well as how each differed across writings. Part 2 looked at collocations and n-grams to determine relevant themes as well as the extent to which various labels are used to describe Autistic individuals in each community.

2 Procedure

We compiled a corpus of web-based writings on the subject of vaccination and annotated for two classes: pro-vaccination and anti-vaccination. We started with basic corpus analysis, such as keywords, collocations, and n-grams. Then we explored automatic approaches to distinguish between anti- and pro-vaccination texts using psychometric word analysis with LIWC [23] combined with feature selection and classification tasks.

Corpus Collection. A master corpus of 124 separate documents was created primarily focused on the subject of vaccination, with a secondary focus of autism. The corpus was compiled by searching the terms vaccination, vaccinate, vaccine(s), vaccine dangers, anti-vaccine, autism vaccine, vaccine risks on the CDC webpage, EBSCOhost, Generation Rescue, Google, Google Scholar, WHO, and on a variety of news websites purpose chosen to span a variety of political ideologies (e.g. ABC, BBC, Breitbart, CNN, Fox). The writings were converted into plain text files, cleaned by hand, and standardized to American English (to improve corpus analysis results).

A human annotator proofread the corpus and eliminated irrelevant material. Our final corpus contains 59 anti-vaccination documents with 92,866 total work tokens and 9,446 word types and 65 pro-vaccination documents with 84,443 word tokens and 8,307 word types.

2.1 Vaccination Sentiment Annotation

Each piece of writing was read and categorized as either pro-vaccination: encouraging vaccination efforts and/or discussing the benefits of vaccination, or anti-vaccination: discouraging vaccination and/or discussing the risks. Writings that did not seem to fit either category were discarded. Documents were then sorted into one of five genres: news, testimonial, blog, information, or research study. These five categories were created to ensure that each class was composed of a balanced representation of writing styles so it could be safely compared.

The categories were assigned as follows. News was assigned if a piece of writing was an impersonal (to the author) story about an event or series of events. A testimonial was assigned if a piece of writing illustrated a personal experience or the experience of a person with a close relationship to the author. Blog was assigned to writings with a clear position written as a personal opinion or the opinion of an organization or cause. Information was assigned if a document shared information proven to be factual or believed to be factual and/or methods for accomplishing something. Research study was assigned if an academic writing was based around an experiment, literature review, or policy.

Corpus Analysis. Collocations (within five words to the left or right: 5L, 5R) were found in each corpus surrounding the terms: autism, autistics, had autism, has autism, is autistic, with autism, child, children, parent, parents, person, people, vaccine, vaccines Collocates were found using (MI) Mutual Information and T-Score [24]. Common n-grams (trigrams, 4-grams, and 5-grams) were also extracted.

We used part of a general-purpose corpus of contemporary American English (COCA) [25] to define keywords. Spoken texts were removed in order to reflect the genre variation of our corpora, which did not include spoken items. Keywords and the terms from each class with a relative frequency over 4 were tested for significance using tf-idf weights.

2.2 Feature Selection

We extracted 95 features using LIWC2015 [23]. Linguistic Inquiry and Word Count (LIWC) is an "efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' written speech samples" [26]. By using LIWC features, we endeavor to explain and compare motivations behind the language use of pro-vaccination and anti-vaccination advocates.

We applied six different feature selection (Information Gain, Cfs Subset, Wrapper Subset, Chi Squared, Gain Ratio, Symmetrical Uncertainty) algorithms to our data and chose only those on which all six agreed. These 16 LIWC features, Words/sentence, 1st person singular, 3rd person singular, Cognitive processes, Insight, Differentiation, Body, Relativity, Space, Swear words, Periods, Commas, Question marks, and Exclamation marks, were used for the majority of our classification experiments.

Table 1. Classification task performance.

features	Model	Accuracy	Precision	Recall	F1
ALL (95)	Baseline	**0.52**	0.28	0.36	**0.47**
	NaivesBayes	0.69	0.69	0.69	0.68
16	J48	0.68	0.68	0.68	0.68
	DecisonTable	0.69	0.69	0.69	0.69
	NaïveBayes	**0.77**	0.80	0.77	**0.77**

Tasks with other classification algorithms were attempted and yielded less notable results.

2.3 Classification

Due to our balanced corpus, we chose a simple majority class classifier for our baseline. We also experimented with other classification algorithms and with different feature sets. Our results are summarized in Table 1.

The Naive Bayes classifier produced promising results (77%) and significantly improved over the baseline (52%) when used with the 16 features described above and 10-fold cross-validation.

Table 2. Keywords by class.

rank	anti-vaccination	pro-vaccination
1	vaccine	vaccine
2	vaccines	vaccines
3	autism	vaccination
4	vaccination	measles
5	vaccinations	autism
6	immune	vaccinated
7	mercury	immunity
8	vaccinated	rubella
9	thimerosal	mmr

3 Results

The most frequent words across corpora were similar. As seen in Table 2, 11 of the top 20 words are shared between the pro- and anti-vaccination categories. However, tf-idf weights revealed fewer similarities across categories (Table 3).

Looking at tf-idf weights, *DNA* has the most weight in the anti- vaccination corpus. This aligns with the mistaken belief in the anti-vaccination community that vaccines made using human embryo cells can cause harm to patients [27]. This term along with *flu* highlights the prevalence of misinformation beyond the vaccine-autism link. Whereas *thimerosal* (a mercury-containing organic compound used as a preservative in some vaccines), *mercury*, and *autism*, which also show up in the keywords and most frequent words, are clear references to the beliefs brought on by Wakefield [5]. The high rank of *I* and *he* are in line with the LIWC feature categories 1st person singular, and 3rd person singular in this corpus, which through multiple feature selection and classification tasks were determined to be significant.

Table 3. tf-idf weight by class.

rank	anti-vaccination		pro-vaccination	
1	dna	11.16	sliv	27.78
2	i	7.54	consent	13.59
3	thimerosal	7.49	influenza	11.50
4	pregnancy	7.37	herd	10.36
5	mercury	6.86	exemption	9.44
6	flu	6.47	hpv	9.03
7	cell	6.32	schools	8.50
8	he	6.06	exemptions	8.38
9	cells	5.86	immunity	8.21
10	autism	5.79	participants	7.80

sliv refers to school-located influenza vaccination

In the pro-vaccination corpus the highest tf-idf weights were *sliv (*school-located influenza vaccination), *consent*, *influenza*, and *herd*. These terms highlight the policy efforts of pro-vaccination proponents to lower the age of consent for vaccination. Vaccination in schools leverages higher coverage to improve herd immunity, as school children have the highest rates of infection. These terms along with *schools* and *immunity*, show that vaccine coverage in schools and herd immunity awareness are issues central to vaccination advocates' efforts.

Table 4. Most frequent words by class.

	instances of word in corpus (rank)	
word	**pro-vaccination**	**anti-vaccination**
vaccine	905 (1)	1001 (1)
vaccination	626 (2)	225 (6)
vaccines	559 (3)	671 (2)
children	475 (4)	466 (3)
measles	454 (5)	---
autism	270 (8)	378 (4)
health	286 (7)	279 (5)
parents	312 (6)	184 (9)
medical	---	199 (7)
child	---	195 (8)
people	260 (9)	130 (18)
disease	206 (13)	162 (10)
vaccinated	220 (10)	---

4 Discussion

The high term frequency words, keywords, and tf-idf representations in our pro-vaccination corpus are in line with the results of Kang et al. [17] with respect to pro-vaccination writings. Their positive network included the concepts: measles, autism, HPV vaccine, vaccine-autism link, meningococcal disease, and MMR vaccine. Their negative network referred to the CDC, vaccine industry, doctors, mainstream media, pharmaceutical companies, and United States. While the frequent words (Table 4), keywords, and tf-idf in our anti-vaccination corpus were not consistent with these findings, as they included more language focused on vaccine ingredients and the pur-

ported negative effects of vaccination, the n-grams revealed terms in line with these concepts and agencies.

However, the pro-vaccination n-grams also largely referenced the United States and CDC as well. There was a notable lack of mentions of pharmaceutical companies across both corpora. Possible reasons for these differences could be the addition of autism as a search term during our corpus compilation, and the exclusive use of news articles in Kang et al.'s study contrasted with our inclusion of other genres.

As far as punctuation, the anti-vaccination texts contained more question marks and exclamation marks than the pro-vaccination texts (see Fig. 1). The use of exclamation marks in anti-vaccination texts was also notably ten times the amount of baseline LIWC tests [26]. This is indicative of stylistic differences that may be the result of the prevalence of funded non-profit organizations' writings in the pro vaccination corpus. Such organizations would have the resources for widespread quality control (i.e. proofreading). In contrast, most anti vaccination writing were published by individuals or online-community interest groups that, with the exception of Generation Rescue, would not have such resources. Informality was also evidenced by more than six times the amount of swear words in the anti class than the pro. The combined vaccination corpus, however, contained far less profane words than baseline. Pro-vaccination writings also contained more periods and commas than anti vaccination writings. The role of punctuation may be related. The accessibility of resources mentioned above as well as the tendency for academic writing to contain more periods and commas than are typically found in less formal styles due to in text citations.

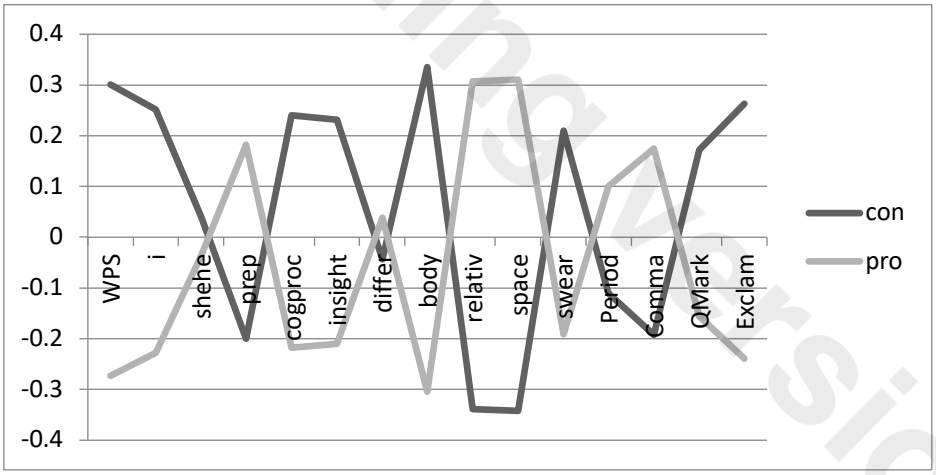

Fig. 1. Normally distributed values across LIWC features of interest for pro-vaccination and anti-vaccination classes.

Pro-vaccination writings made more references to relativity and space (Fig. 1), perhaps a result of advocates' reporting on regions of outbreak, spread of disease, and resurgences as well as the tendency to explore best practices for vaccination coverage which include location and strategy. The anti-vaccination class made fewer references

to words in these categories, and far less than baseline, indicating that perhaps location and movement is less important to their cause.

Anti-vaccination texts contained twice the mentions of body-associated words when compared to pro, which surprisingly contained less than baseline. These findings agree with those of Faasse and Chapman [3], which describe anti-vaccine Facebook comments as having more body references than pro-vaccine comments. Past research has shown that the typical person with anti-vaccination attitudes is more disgusted by blood and needles than a person without such beliefs [12]. Perhaps a larger percentage of body related words reflect these fears. Anti-vaccination results also support findings from that anti-vaccine comments make more references to money [3]. Pro-vaccination results run counter to Faasse and Chapman [3], as anti-vaccination writings made slightly more references to family, though both contained a much higher percentage of family words than baseline.

Cognitive process words ranked higher, in anti-vaccination writings, as well as its subcategory insight. Use of insight words while detailing past events implies the processing of trauma [28]. One reason there may be a larger percentage of such words is that many anti-vaccination advocates are grieving. The term SIDS (sudden infant death syndrome) ranked high (14th of anti 4-grams, and a frequency of 23) in several metrics. These LIWC insight percentages may be indicative of a grieving process that many parents are going through. Some parents even consider having an autistic child as a loss. In a 2010 interview with PBS Frontline, Jenny McCarthy, anti-vaccination advocate, is quoted as saying, “Ask 99.9 percent of parents who have children with autism if we'd rather have the measles versus autism, we'd sign up for the measles.” Whether overt, or subtler, the grieving process is a possible reason for a larger percentage of cognitive process words. In pro-vaccination writings where it is less common to see themes of trauma and loss, these cognitive process words are less prevalent. A heightened percentage of cognitive processes terms is also a quality of more complex language, but so are prepositions of which the pro-vaccination writings contained slightly more [29].

Table 5. Label collocations across classes.

phrase	has autism	had autism	is autistic	with autism	autism	autistic
anti collocates	skeparent	exhausting	eldest	children	spectrum	children
	blogher	son	complications	leucocytes	adhd	
	vaccinate	mccarthy	son	exploration	relationship	
	medicine	jenny	autistic	anomalies	prevalence	
	based	never	reasons	accommodations	disorder	
pro collocates		result		toddler	trigger	undisclosed
		never		subsets	spectrum	transpired
		three		spurious	disorder	integrating
		being		rarer	cause	hating
		vaccinated		proportionally	link	faces

Blank space indicates a lack of statistically significant collocates outside of function words.

Anti-vaccination writings contained more 1st person pronouns than pro-vaccination and baseline. These findings are consistent with Kowalski [30] who found

that victim-focused language used more 1st person singular pronouns than language where the writer is at fault. This is consistent with the "vaccines as the villain" narrative found in many anti-vaccination writings. Greater use of 1st person also corresponds to emotion and informal language and 3rd person to social interests and support [26]. Considering that the majority of anti-vaccination writings came from sources with less resources, it may be the case that they would contain more emotional and less formal language, and read as if from a social interest or support group.

Though a larger corpus is needed to comprehensively evaluate the impact of labels in vaccination writings, there is evidence of a higher percentage of noun labels within anti-vaccination writings when compared with pro-vaccination sources (Table 5). While any label can be harmful, noun labels have been shown to most negatively influence others' perceptions of neuro-divergent individuals [20].

Conclusion

Overall, the main divisions that categorize anti-vaccination writings from pro-vaccination tend to concern formality and narrative structure: mostly notably use of punctuation, pronouns, and cognitive process words. Themes and key terms were concurrent to past research with one exception. However, more research is needed to see if individual genres within the subject of vaccination have unique qualities. By compiling a larger corpus, future studies could explore differences across categories of writing, and perhaps uncover further relevant LIWC features. Evidence suggests that future health promotion efforts should make attempts not to talk down to individuals, given the writings' less academic qualities and should stem from a place of understanding, considering the grieving process many anti-vaccination advocates may be experiencing.

Waning vaccination coverage and recent resurgences of once "eradicated" diseases have drawn the attention of public health officials, policy makers, and governments worldwide. As anti-vaccination sentiment continues to flourish, understanding the attitudes of its supporters and the nuances of the movement's misinformation become an essential to health promotion efforts. Further research that expands upon this pilot study is cogent to these efforts. Anti-vaccination attitudes have a long history, but with new strategies and new information, research of this kind can only serve to enrich the fight for maximum vaccination coverage, herd immunity, and a healthier world.

Acknowledgments. Thank you to Lauren Covey, Anna Feldman, and Susana Sotillo for teaching me and supporting this process. R. Claire Lingenfelter, thank you for the moral support and Vikki Fiore for the moral and technical support.

References

1. Ten Threats to Global Health in 2019. World Health Organization website. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019. Updated January 2019. Accessed March 3, 2019.

2. Jolley D, Douglas KM. The effects of anti-vaccine conspiracy theories on vaccination intentions. PloS one. 2014;9(2):e89177.

3. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro-and anti-vaccination comments in response to a high profile Facebook post. Vaccine. 2016;34(47):5808-14.

4. Schmidt AL, Zollo F, Scala A, Betsch C, Quattrociocchi W. Polarization of the vaccination debate on Facebook. Vaccine. 2018;36(25):3606-12.

5. Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet.* 1998;352(9123):234-5

6. Monitoring Measles Cases. Centers for Disease Control website. https://www.infectioncontroltoday.com/public-health/cdc-monitoring-measles-cases. Updated August 17, 2018. Accessed March 3, 2019.

7. Caplan AL. Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Weekly Epidemiological Record.* 2009;84:301-8.

8. Hviid A, Hansen JV, Frisch M, Melbye M. Measles, mumps, rubella vaccination and autism: A nationwide cohort study. *Annals of Internal Medicine*. 2019;170(8):513-20.

9. Madsen KM, Hviid A, Vestergaard M, Schendel D, Wohlfahrt J, Thorsen P, et al. A population-based study of measles, mumps, and rubella vaccination and autism. *New England Journal of Medicine*. 2002;347(19):1477-82.

10. Stratton K, Gable A, Shetty P, McCormick M, Institute of Medicine (US) Immunization Safety Review Committee. Immunization safety review: measles-mumps-rubella vaccine and autism. *National Academies Press*. 2001.

11. Taylor B, Miller E, Farrington C, Petropoulos MC, Favot-Mayaud I, Li J, et al. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *The Lancet.* 1999;353(9169):2026-9.

12. Hornsey MJ, Harris EA, Fielding KS. The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology*. 2018;37(4):307.

13. Backus F, De Pinto J, Dutton S, Salvanto A. CBS News poll on vaccines and the measles outbreak. *CBS News*. February 20, 2015 https://www.cbsnews.com/news/cbs-news-poll-on-vaccines-and-the-measles-outbreak. Accessed March 3, 2019.

14. Hill HA, Elam-Evans LD, Yankey D, Singleton JA, Kang Y. Vaccination coverage among children aged 19–35 months—United States, 2016. *MMWR*. Morbidity and mortality weekly report. 2017;66(43):1171.

15. Horne Z, Powell D, Hummel JE, Holyoak KJ. Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences.* 2015;112(33):10321-4.

16. Joslyn MR, Sylvester SM. The determinants and consequences of accurate beliefs about childhood vaccinations. *American Politics Research.* 2019;47(3):628-49.

17. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. *Vaccine.* 2017;35(29):3621-38.

18. Arif N, Al-Jefri M, Bizzi IH, Perano GB, Goldman M, Haq I, et al. Fake news or weak science? Visibility and characterization of anti-vaccine webpages returned by Google in different languages and countries. *Frontiers in Immunology.* 2018;9:1215.

19. Nyhan B, Reifler J, Richey S, Freed GL. Effective messages in vaccine promotion: a randomized trial. *Pediatrics.* 2014;133(4):e835-42.

20. Cuttler C, Ryckman M. Don't call me delusional: Stigmatizing effects of noun labels on people with mental disorders. *Stigma and Health.* 2019;4(2):118.

21. Brown L. Identity-First Language. *Autistic Self Advocacy Network.* March 2, 2012. https://autisticadvocacy.org/about-asan/identity-first-language. Accessed April 10, 2019.

22. Zerbo O, Modaressi S, Goddard K, Lewis E, Fireman BH, Daley MF, et al. Vaccination patterns in children after autism spectrum disorder diagnosis and in their younger siblings. *JAMA pediatrics.* 2018;172(5):469-75.

23. Linguistic Inquiry and Word Count. LIWC2015. Austin,TX: Pennebaker Conglomerates; 2015.

24. Stubbs M. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language.* 1995;2(1):23-55.

25. Davies M. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*. 2010;25(4):447-64.

26. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. 2015.

27. Yang H. Establishing acceptable limits of residual DNA. *PDA Journal of Pharmaceutical Science and Technology.* 2013;67(2):155-63.

28. Pennebaker JW, Mayne TJ, Francis ME. Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology.* 1997;72(4):863.

29. Hartley J, Pennebaker JW, Fox C. Abstracts, introductions and discussions: How far do they differ in style?. *Scientometrics.* 2003;57(3):389-98.

30. Kowalski RM. "I was only kidding!": Victims' and perpetrators' perceptions of teasing. *Personality and Social Psychology Bulletin.* 2000;26(2):231-41.